DOCUMENT RESUME

ED 353 840 FL 020 977

AUTHOR Hill, David J.
TITLE Cluster Analysis and the Interlanguage Lexicon.
REPORT NO ISSN-0959-2253
PUB DATE 92
NOTE 13p.; For serial issue in which this paper appears, see FL 020 971.
PUB TYPE Reports - Research/Technical (143) -- Journal Articles (080)
JOURNAL CIT Edinburgh Working Papers in Linguistics; n3 p67-77 1992

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Cluster Analysis; *English (Second Language); Foreign Countries; *Interlanguage; Language Research; Multidimensional Scaling; Native Speakers; *Task Analysis; Verbs; *Vocabulary
IDENTIFIERS *Kenya

ABSTRACT

As part of a research program investigating the interlanguage lexicon of Kenyan learners of English, based on the lexical field of locomotion, a card sorting task was used to produce raw similarity measure data. This was subjected to cluster analysis and multidimensional scaling (MDS). This paper suggests that cluster analysis/MDS of perceived similarities of selected verbs in a sorting task may indicate possible lines of investigation of the lexical knowledge of different groups of learners in comparison to that of native speakers. (Contains 25 references.) (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED353840

Cluster Analysis and the Interlanguage Lexicon

David J. Hill (DAL)

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Brian Parkinson

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)"

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it

☐ Minor changes have been made to improve reproduction quality

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

FL 020 #08

BEST COPY AVAILABLE

CLUSTER ANALYSIS AND THE INTERLANGUAGE LEXICON

David J. Hill (DAL)

Abstract

As part of a research programme investigating the interlanguage lexicon of Kenyan learners of English, based on the lexical field of locomotion, a card sorting task was used to produce raw similarity measure data. This was subjected to cluster analysis and multidimensional scaling (MDS). This paper suggests that cluster analysis/MDS of perceived similarities of selected verbs in a sorting task may indicate possible lines of investigation of the lexical knowledge of different groups of learners in comparison to that of native speakers.

1. Research context

This analysis was carried out as part of my research into the interlanguage (IL) lexicon of L2 learners of English in Kenya. By IL lexicon is here meant the lexical knowledge of the target language displayed by a second/foreign language learner. It includes both what the learner uses productively (which is directly observable) and also what he uses receptively (which is only indirectly observable). This working definition takes no position regarding the psycholinguistic nature of this lexical knowledge or its relation to native or other language knowledge. The research objective was to see if there were any significant differences in the lexical performance of learners of English from different first language backgrounds.

The research was based on the lexical field of locomotion, which has been investigated in some depth by linguists of different persuasions (see for example Leech 1969, Ikegami 1970, Miller 1972, Lyons 1977, Talmy 1975, 1985). Two basic trends can be discerned in the literature. Motion is analysed in one as a transition between locational states, while in the other motion itself is seen as fundamental, along with such notions as source, path and goal. The latter theoretical framework has been adopted here and use has been made of Talmy's distinction between verbs that incorporate path and motion and those that incorporate manner and motion. English appears to belong to a group of languages that make particular use of manner verbs of motion (Talmy 1985).

The results of a story retelling experiment indicated a strong overall preference on the part of the Kenyan learners for path-specifying motion verbs, with possible differences related to mother tongue (see Hill 1991). Three mother tongues were represented among the subjects who performed the task discussed in this paper: Luo (32 speakers), Nandi (35) and Olunyore (26). The first two are Nilotic languages and therefore distantly related; the last is a Bantu language. The subjects were all trainees at two primary teachers colleges in Western Kenya and had had at least ten years' education in English. Native speaker benchmarks were provided by 31 children at two secondary schools in the Edinburgh area.

EDINBURGH WORKING PAPERS IN APPLIED LINGUISTICS, number 3 (1992) ISSN 0959-2253

2. The card-sorting technique

Psychologists have used sorting - also known as direct grouping - as one of several methods to investigate the mental lexicon. Typically, subjects are given a set of cards with words or sentences printed on them and are asked to sort them according to similarity of meaning into as many classes as they wish. The method has been tried with different classes of words. Clark (1968) used it in a study of English prepositions. Miller (1969) tested his subjects with a heterogeneous collection of 48 nouns. Fillenbaum and Rapoport (1971) looked at words in a number of lexical fields, such as colour terms, verbs of possession and evaluative adjectives. Kellerman (1978) employed the method with a set of sentences using a single Dutch verb in a variety of literal and metaphorical senses.

The theory behind each experiment has depended on the preferences of the researcher. Miller (1969) made the assumption that native speakers would sort nouns according to the semantic features they share while overlooking their distinguishing features. By pooling the data from a group of subjects a measure of similarity of two items could be obtained from the number of times the two items were put together in the same pile. The higher the number, the greater the degree of adjudged similarity for that group.

However, feature theory is no longer as fashionable as it was when Miller wrote his paper. It has come under attack from various quarters - theoretical linguistics (e.g. Bolinger 1965, Lyons 1977, Sampson 1979), anthropology (e.g. Tyler 1978) and, more recently, cognitive linguistics (Langacker 1987, Lakoff 1987). It has been criticized for its reductionist 'atomistic' approach to meaning, which imposes an arbitrary structure in which there is no self-evident way of showing which senses are more important than others. There is also no theoretical limit to the number of features that can be identified and we lack a metalanguage to describe some of the components. Nevertheless, it seems difficult to carry out any kind of contrastive lexical analysis without making use of some kind of semantic features - 'as problem-ridden in theory as [they are] indispensable in practice' (Cruse 1986: 22; he prefers the term 'semantic trait').

However, cognitive semantics may provide a solution in the form of prototype theory. To take a relevant example, each of us has a mental image of a prototypical act of running; though this image may differ in some details from person to person and between cultures, a prototypical image is essentially unfocused and non-specific (cf. Lakoff 1987). Now it could be claimed that in a sorting task we are in effect being asked to compare our mental images associated with a number of lexical items and that this is something we do holistically and not on a point by point basis according to what Fillmore (1975) calls a checklist theory of meaning. This does not preclude us from subsequently trying to justify our sorting on the basis of particular shared characteristics.

All of the studies mentioned were concerned with the judgments of native speakers. The present researcher considered that it would be interesting to see what results could be obtained using the card sorting technique with groups of learners from different language backgrounds.

The learners were, as already indicated, at a fairly advanced level; they were undergoing training to teach in schools using English as the medium of instruction. As for the lexical items, the majority (13 out of 20) are to be found in the General Service List. Four of the remaining items are in the Cambridge English Lexicon (i.e. within the

comprehension level of the Cambridge First Certificate). That leaves three items - STROLL, STAGGER and TIPTOE - which might be considered less common, though they are in the Longman English Lexicon, i.e. words that learners at this level would be expected to have at least receptive knowledge of. The full list of items appears below in Table 1.

All the items were inserted in a common sentence frame in order to provide at least a minimal meaningful context:

Juma [V-ED] along the path.

The items themselves were given prominence by being typed in upper case. Each sentence was presented on a card measuring 6cm by 11cm. Individual subjects were handed a set of cards and asked to arrange them into groups of similar meaning. They were told they could make as many groups as they wished and put any number of cards in each group. Apart from this they were not given any further explanation of what was meant by similarity of meaning. Nearly everybody set about the task quite readily. Subjects worked independently but were often in the same room as others doing the same or other tasks. They took between three and ten minutes to make their groupings. No time limit or other constraint was imposed on them. Some agonized for long periods over the allocation of one or two cards or, in rare cases, made complete rearrangements. At the end a few indicated that they were not completely happy with their final choice or pointed out that other orderings would have been equally good. The number of piles made ranged from 3 to 15, with an overall mean of 7.96 and a standard deviation of 2.54; the small differences between the group means were not statistically significant.

3. Cluster analysis

The similarity measures for each group of subjects can be shown in a half matrix. Table 1 gives the raw data for the native speakers. This shows that, for example, 26 subjects sorted LIMP and STAGGER together, whereas only 3 put COME and WANDER in the same pile. It is possible to make interpretations directly from this data, but cluster analysis enables this to be done in a more orderly way.

The origins of this type of analysis can be found in the development of methods of numerical taxonomy in botany and zoology which started in the late eighteenth century. Cluster analysis is itself the name for a cluster of techniques - mostly formulated by mathematical statisticians in the late 1960s and early 1970s as computers came to be more widely used - for grouping together entities of any sort (cf. Everitt 1974; Lorr 1983; Aldenderfer and Blashfield 1984; Kaufman and Rousseeuw 1990). The basic idea is that entities (which can be individuals or objects) in the same cluster are more like each other than they are like those in other clusters. The techniques of cluster analysis are therefore designed to maximize differences between clusters relative to variation within clusters. Similarity or dissimilarity measures (such as those in Table 1) can also be regarded as distances between entities. There are two general categories of clustering algorithms, hierarchical and partitioning. Hierarchical procedures, which involve the construction of tree-like structures, can be either divisive or agglomerative. We shall only be concerned here with the last type.

Table 1: Similarity matrix for native speaker group

```
A advanced
5  B came
0  1  C crept
3  0  0  D dashed
7  1  0  9  E escaped
5  0  0 23 10  F hurried
3  2  1 16  7 15  G jogged
2  0  8  1  1  1  2  H limped
8  4  2  5  4  4  7  1  J marched
9 10  1  3  3  2  4  2  5  K moved
11  9  1  2  3  2  3  3  4 11  L passed
4  0  0 25  8 25 16  0  3  2  1  M ran
6  8  0  2 10  2  1  1  1  2  2  1  N returned
1  2  6  1  1  1  2 26  2  3  2  0  2  P staggered
5  6  7  2  1  2  4  6  7  8  7  1  3  8  R strolled
1  3 23  1  1  1  2  8  2  4  1  0  2  9 10  S tiptoed
13  4  1  5  6  5  5  1 10  9  8  4  5  1  9  1  T travelled
8 10  3  1  1  1  3  2  8 15 13  0  2  4 18 10  5  U walked
3  3  4  1  3  1  2  5  7  6  4  0  3  6 15  9  5  9  V wandered
7 20  3  0  2  0  2  1  4 15 13  0  3  1  5  7  3 10  4  W went
```

In agglomerative methods the entities with the shortest distance between them are linked first, then those with the next shortest distance, and so on until all the entities are joined in one all-embracing group. Thus in single linkage - also known as the nearest-neighbour method - groups initially consisting of single individuals are fused according to the distance between their nearest members, the groups with the smallest distance being fused first. Each successive fusion decreases by one the number of groups. The furthest-neighbour or complete linkage method is exactly the opposite in that the distance between groups is now defined as that between their most remote pair of individuals. Average linkage is a compromise between these two methods. Other methods exist, with varying degrees of mathematical complexity, such as Ward's method and the centroid method. The choice of method depends very much on what the investigator is using cluster analysis for, but even so is by no means clear-cut.

It should be apparent that cluster analysis is essentially exploratory and descriptive in nature. Since the investigator is free to choose

- the type of similarity/dissimilarity measure
- the clustering algorithm
- the cut-off point for determining the number of significant clusters (though 'rules' do exist for this)

this gives cluster analysis a great deal of flexibility. It also renders its interpretation somewhat problematic. The need for some kind of validation technique is stressed by several authorities (e.g. Aldenderfer and Blashfield 1984). The only one which it was possible to use here was replication, i.e splitting each sample into two and performing cluster analysis on both to check for internal consistency.

Because of the large amounts of computation involved, cluster analysis is only practical with a computer program. I used the CLUSTER program on the SPSS-X statistical package. This program can produce several kinds of output, but probably the most useful and easiest to understand is the tree graph or dendrogram. Figures 1 and 2 are based on the dendrograms produced for two of the groups using the method of group average linkage, which avoids the extremes of single and complete linkage - these tend to produce chain-like clusters and over-compact clusters respectively - and is recommended for general use by, among others, Kaufman and Rousseeuw (1990).

For the native speakers six clusters of from two to four items stand out quite clearly; these merge together with the outlying items to form three large clusters. A similar statement could be made about the Luo speakers, except that the content of the clusters was in some cases quite different. The grouping of LIMP with STAGGER and subsequently with CREEP and TIPTOE is common to both; indeed these two pairs are clearly seen in all the groups, but do not coalesce in the case of the Nandi speakers. However, for the Luos COME and GO only join up in the larger cluster, while RETURN - an outlier in the native speaker group - is closely linked to COME.

4. Multidimensional scaling

A more graphic way of presenting cluster analysis data is by means of multidimensional scaling (cf. Everitt 1978; Hair et al., 1987). This is a set of techniques related to cluster analysis in that they operate on similarity measures, but the results are obtained by a totally different algorithm.

Multidimensional scaling (MDS) programs attempt to find a set of points in a reduced number of dimensions such that the distances in this lower dimensional space are monotonically related to the similarities in the matrix. The monotonicity property (by which a value either never decreases or never increases) cannot in general be completely satisfied and a measure called 'stress' is used to assess the extent to which a configuration falls short of this requirement. Essentially an MDS program begins from an arbitrary initial configuration and proceeds in a stepwise manner making successive adjustments to the co-ordinates in order to decrease the stress.

Figures 3, 4, 5 and 6 show MDS configurations for each group of subjects. Notice that Figure 6 is in fact a two- dimensional representation of a three-dimensional configuration, which reduced the stress value quite significantly in this case. It should be emphasized that this is simply a way of displaying the data obtained in this experiment. It may be tempting to regard lexical items as floating round in some kind of semantic space in one's head, but an external representation of what someone knows is not necessarily equivalent to the internal form of that person's knowledge (Bransford and Nitsch 1978).

BEST COPY AVAILABLE

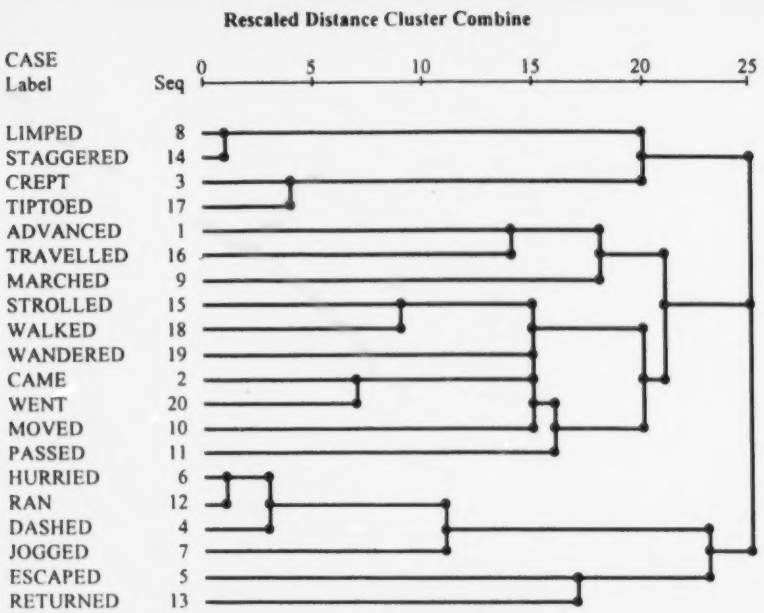

Figure 1: Dendrogram for native speaker group using Average Linkage

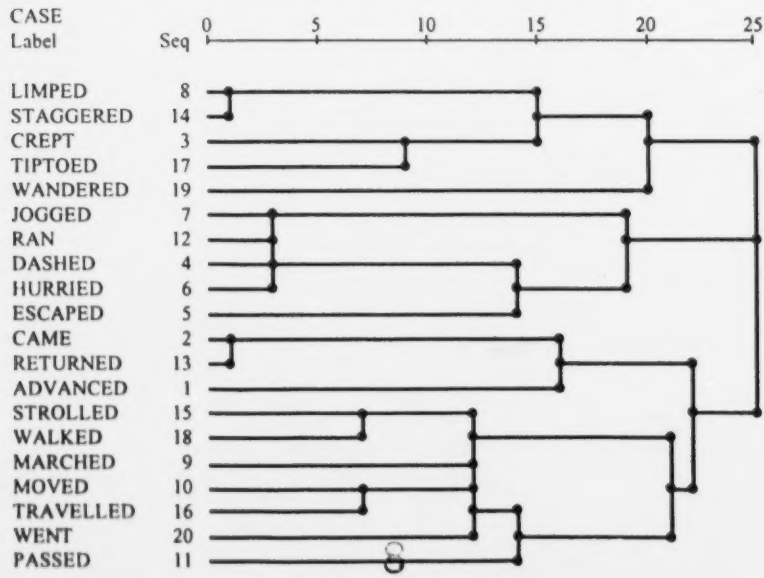

Figure 2: Dendrogram for Luo speaker group using Average Linkage

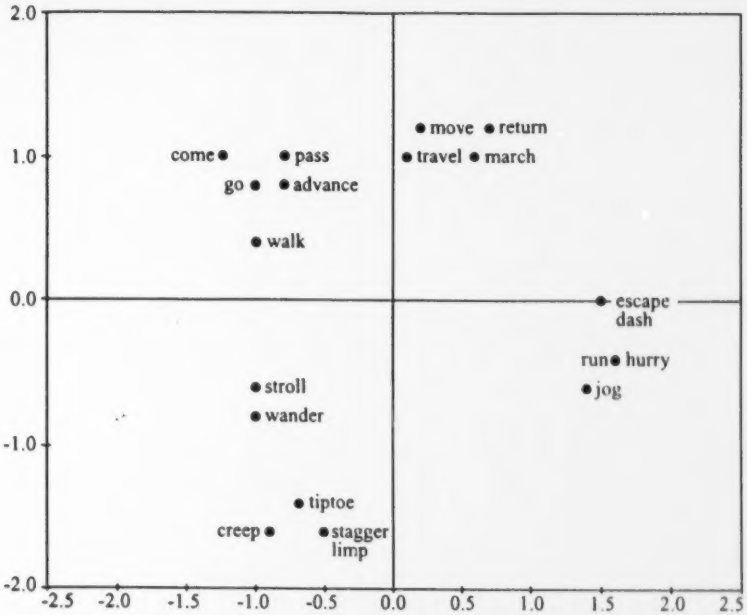

Figure 3: Derived two-dimensional stimulus configuration for native speaker group

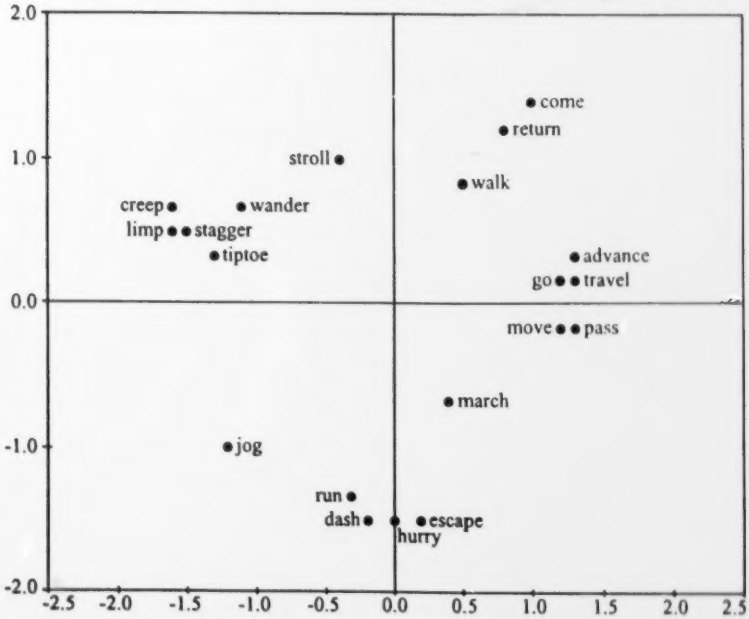

Figure 4: Derived two-dimensional stimulus configuration for Luo speaker group

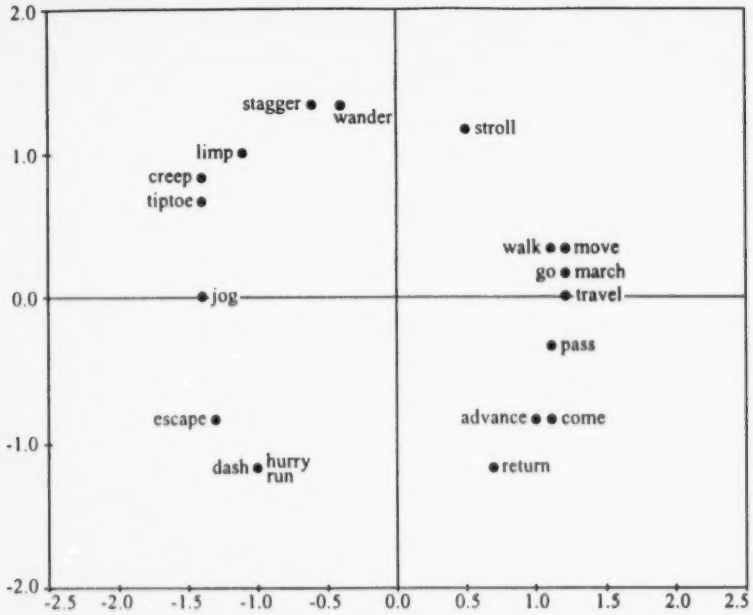

Figure 5: Derived two-dimensional stimulus configuration for Nandi speakers group

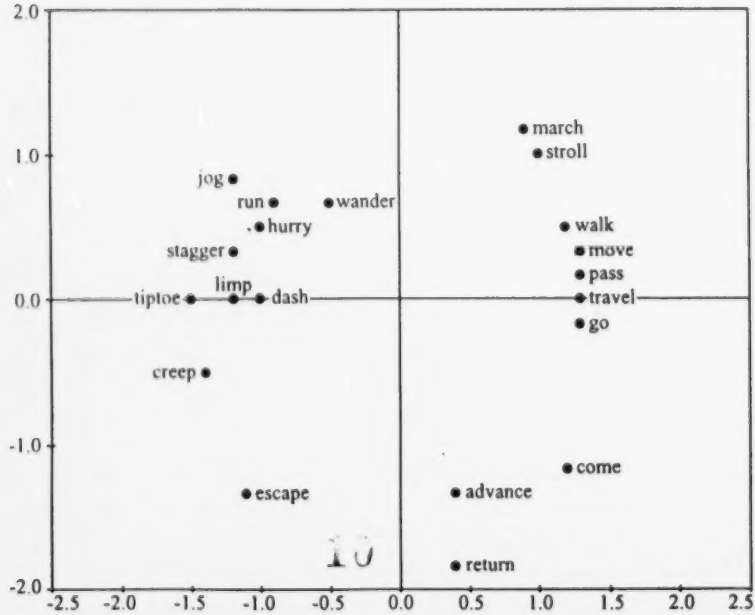

Figure 6: Derived three-dimensional stimulus configuration for Olunyore speaker group

5. Discussion

As suggested earlier, interpretation of cluster analysis and MDS is not a simple matter. In the absence of any real validation tests, one must be very cautious about conclusions. There is also a good deal of 'noise' in the data and there are several intervening variables. Moreover the cognitive demands of this task may not mirror those of 'ordinary' language processing, so we cannot assume that we are gaining insight into actual lexical knowledge. Nevertheless the assignment of COME and GO to separate clusters in all the learner groups seems very clear. COME forms a kind of path/direction group with RETURN and ADVANCE, while GO belongs to a more general group of motion verbs - MOVE, PASS and TRAVEL. For native speakers, COME and GO are much more closely linked; indeed, two-thirds of the sample sorted them together.

There are also differences between the three Kenyan language groups which may be significant. For the two Nilotic language groups ESCAPE tends to go with DASH and HURRY (though less strongly so in the case of the Nandi) to form a group which might be labelled 'rapid movement', while for the Bantu group ESCAPE is linked to CREEP and TIPTOE to make a kind of 'secretive movement' group; this would certainly fit the semantic profile of the corresponding Olunyore verb. ESCAPE is something of an outlier for the native speakers, with weak links to HURRY and RETURN. The separation of LIMP and STAGGER from CREEP and TIPTOE for the Nandi speakers has already been mentioned, although it does not show up so clearly on Figure 5.

A closer examination of Figure 3 might suggest that the vertical dimension could be described as one of 'manner' (all the path verbs are in the top half) and the horizontal dimension as one of 'speed', with RUN and STROLL at the two extremes (though COME and GO should perhaps be nearer the middle than they actually are). The dimension assignments of the other groups are more difficult to determine.

It is obvious that further investigation, using larger samples and a wider variety of techniques, will be needed to test the generalizability of these observations, together with a more profound study of the poorly documented lexicons of the Kenyan languages. However the results so far are at least not inconsistent with the hypothesis that learners' IL lexical categories are influenced by those of their mother tongue.

References

Aldenderfer M.S. and R.K. Blashfield. 1984. Cluster Analysis. Beverley Hills: Sage Publications.

Bolinger D. 1965. 'The atomization of meaning.' Language 41: 555-573.

Bransford J.D. and K.E. Nitsch. 1978. 'Coming to understand things we could not previously understand'. In J.A. Kavanagh and W. Strange (eds) Speech and Language in the Laboratory, School and Clinic Cambridge, Mass.: MIT Press, 267-307.

BEST COPY AVAILABLE 75

Clark H.H. 1968. 'On the use and meaning of prepositions'. Journal of Verbal Learning and Verbal Behavior. 7: 421-431.

Cruse D.A. 1986. Lexical Semantics. Cambridge: Cambridge University Press.

Everitt B. 1974. Cluster Analysis. London: Heinemann Educational Books.

Everitt B. 1978. Graphical Techniques for Multivariate Data. London: Heinemann Educational Books.

Fillenbaum S. and A.Rapoport A. 1971. Structures in the Subjective Lexicon. New York: Academic Press.

Fillmore C.J. 1977. 'Scenes and frames semantics.' In A. Zampolli (ed.) Linguistic Structures Processing. Amsterdam: North Holland, 55-81.

Hair J.F., R.E. Anderson, and R.L. Tatham. 1987. Multivariate Data Analysis (2nd edition). New York : Macmillan.

Hill D.J. 1991. 'Interlanguage lexis: an investigation of verb choice'. Edinburgh Working Papers in Applied Linguistics 2: 24-36.

Ikegami Y. 1970. The Semological Structure of the English Verbs of Motion: a Stratificational Approach. Tokyo: Sanseido.

Kaufman L. and P.J. Rousseeuw. 1990. Finding Groups in Data. An Introduction to Cluster Analysis. New York: Wiley.

Kellerman E. 1978. 'Giving learners a break: native language intuitions as a source of predictions about transferability.' Working Papers in Bilingualism 15: 59-92.

Lakoff G. 1987. Women, Fire, and Dangerous Things. Chicago: Chicago University Press.

Langacker R.W. 1987. Foundations of Cognitive Grammar. Volume I. Stanford, CA: Stanford University Press.

Leech G.N. 1969. Towards a Semantic Description of English. London: Longman.

Lorr, M. (1983). Cluster Analysis for Social Scientists. San Francisco: Jossey-Bass.

Lyons J. 1977. Semantics. Cambridge: Cambridge University Press.

Miller G.A. 1969. 'A psychological method to investigate verbal concepts.' Journal of Mathematical Psychology 6: 169-191.

Miller G.A. 1972. 'English verbs of motion: a case study in semantics and lexical memory.' In A.W.Melton and E. Martin (eds.) Coding Processes in Human Memory. Washington, D.C.: Winston, 335-372.

Sampson G. 1979. 'The indivisibility of words.' Journal of Linguistics 15: 39-47.

Talmy L. 1975. 'Semantics and the syntax of motion.' In J.P. Kimball (ed.) Syntax and Semantics Volume 4 New York: Academic Press, 181-238.

Talmy L. 1985. 'Lexicalization patterns: semantic structure in lexical forms.' In T. Shopen (ed.) Language Typology and Syntactic Description, Volume III, Grammatical Categories and the Lexicon Cambridge: Cambridge University Press, 57-149.

Tyler S.A. 1978. The Said and the Unsaid: Mind, Meaning and Culture. New York: Academic Press.